N94-33811

Performance Measurements and Operational Characteristics of the Storage Tek ACS 4400 Tape Library with the Cray Y-MP EL

Gary Hull
Hughes STX Corporation
4400 Forbes Blvd.
Lanham, MD 20706
ghull@flyfish.stx.com

Sanjay Ranade
Infotech SA Inc,
12303 Sandy Point Court
Silver Spring, MD 20904
infotech@access.digex.com

Abstract

With over 5000 units sold, the Storage Tek Automated Cartridge System (ACS) 4400 tape library is currently the most popular large automated tape library. Based on 3480/90 tape technology, the library is used as the migration device ("nearline" storage) in high-performance mass storage systems. In its maximum configuration, one ACS 4400 tape library houses sixteen 3480/3490 tape drives and is capable of holding approximately 6000 cartridge tapes. The maximum storage capacity of one library using 3480 tapes is 1.2 TB and the advertised aggregate I/O rate is about 24 MB/s.

This paper reports on an extensive set of tests designed to accurately assess the performance capabilities and operational characteristics of one STK ACS 4400 tape library holding approximately 5200 cartridge tapes and configured with eight 3480 tape drives. A Cray Y-MP EL2-256 was configured as its host machine. More than 40,000 tape jobs were run in a variety of conditions to gather data in the areas of channel speed characteristics, robotics motion, timed tape mounts and timed tape reads and writes.

Background

The major objectives of this study, part of the High-Performance Computing and Communications Project (HPCC), were as follows:

1. To establish a set of tape I/O performance measurements associated with the current Y-MP EL hardware configuration for comparison with future technology. The STK ACS 4400 Tape Library is the first magnetic tape library system available to the project.

2. To utilize the Cray Y-MP EL as a research tool dedicated to I/O performance measurements.

3. To apply the results of this research to the user community.

Test Environment

This section discusses the computer, disks, tapes, library hardware, software and the system configuration used for the STK ACS 4400 tape library performance tests.

a. Hardware Configuration

The hardware configuration for these tests is shown in Figure 1. The Cray Y-MP EL is a two-processor machine configured with 256 MBytes of central memory. Connected to the main

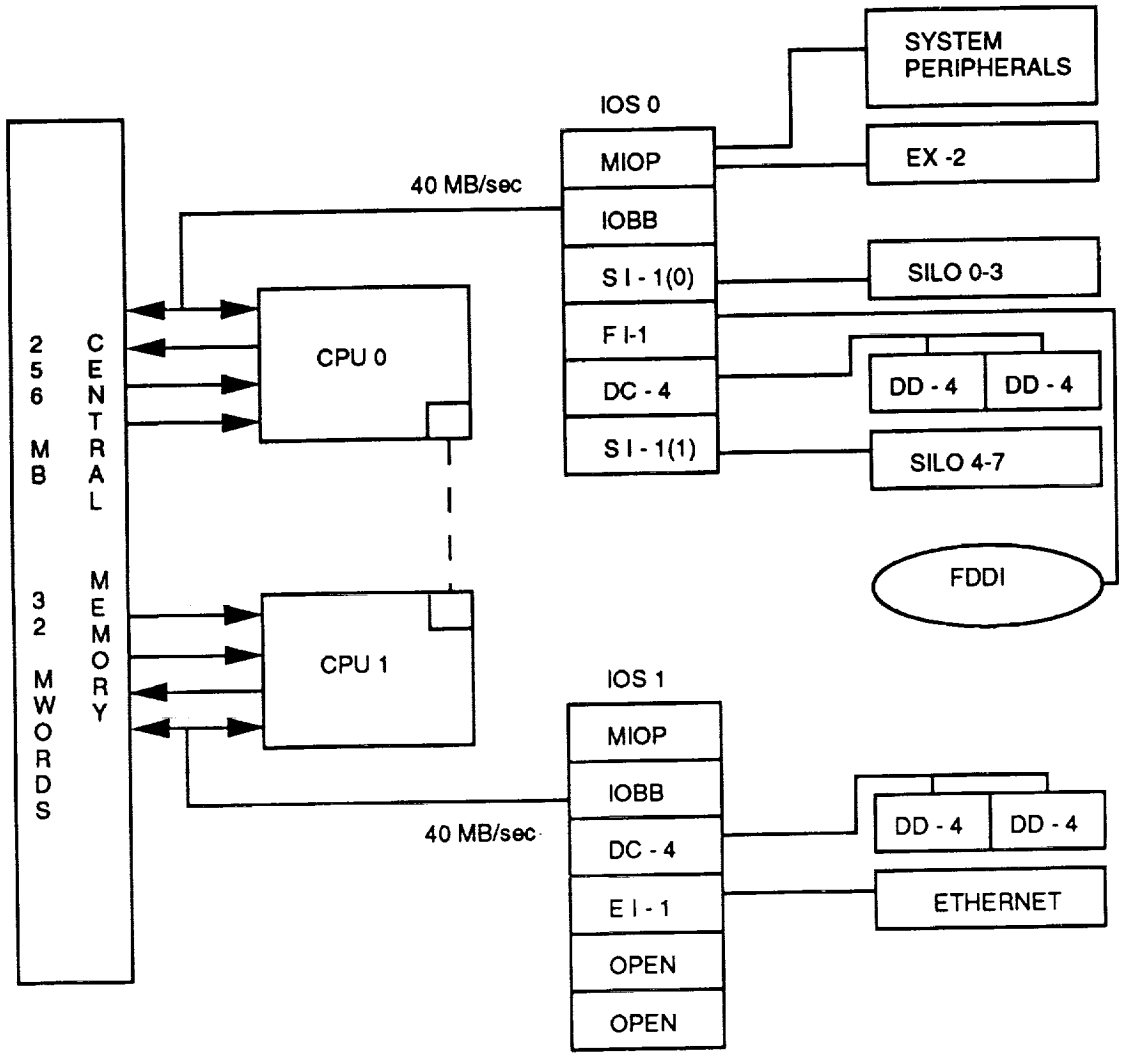

Figure 1. NASA/GSFC Y-MP EL/2-256

memory via two 40 MByte/sec channels are two Input/Output Buffer Boards (IOBB). Connected to the disk controller DC-4 are four DD-4 disks, each with 2.7 GB formatted capacity. The specified peak transfer rate for the DD-4 is 7.5 MByte/sec. These disks are distributed across two controllers and each controller is connected to its own Input -Output Subsystem (IOS).

The tape drives in the STK ACS 4400 library are connected to the CRAY Y-MP EL by two Ciprico SCSI interfaces, each capable of sustaining an advertised transfer rate of 4.5 MByte/s. Both 4781controllers are connected to IOS 0. Each Ciprico controller manages four STK tape drives. The data buffer size for the tape controller is large, but nevertheless limited by the 128K maximum allowed by the IOBB. (Note that the data transfer rate obtainable with the STK ACS 4400 SCSI tape drives is dependent on the specific SCSI interface used to connect the host to the drive. The Ciprico SCSI interface does not have the SCSI incompatibility problem identified with other SCSI interfaces).

The Cray Y-MP EL shared the same Ethernet rib as the STK Sun 330 server and the two systems communicated with each other using standard TCP/IP protocol.

The FDDI connection links the Cray Y-MP EL to the NASA/Goddard campus-wide network.

b. Software Configuration

Release level 6.1.6 of the UNICOS operating system was run on the Cray Y-MP EL and included Cray proprietary software subsystems: Cray tpdaemon, stknet and Data Migration Facility (DMF). Stknet is the software interface which communicates directly with the ACSLS Client Server Interface (CSI) running on the STK Sun 330 server. The ACSLS server software was run at level 3.0. Tape requests from the Cray Y-MP EL are initiated and managed by the Cray tpdaemon, forwarded to stknet and processed as level 3 TCP/IP packets. These packets are then sent to CSI on the STK Sun 330 server using standard TCP/IP protocol (see Figure 2).

The software subsystem DMF manages on-line mass storage space and implements data retrieval and storage to and from the tape library. DMF requests are initiated and managed by the dmdaemon and utilize the same transfer path described above.

c. Test Methodology

More than one hundred programs written in C and Bourne shell scripts were designed and implemented to gather systems and user performance characteristics of the Cray Y-MP EL using the STK 4400 tape library. Emphasis was placed on learning what the general user might experience so accurate predictions could be developed. Snapshots of the Cray Y-MP EL resource allocation characteristics were captured during individual components of each timed test. The data for CPU time, user time and system idle time were used to predict tape I/O resource requirements.

Actual mount times, data write and data read rates were timed under various system load conditions and block sizes. Mount time is defined as the time from the initial mount request by the Cray Y-MP EL to when the STK Sun 330 server replies , acknowledging that the requested tape volume is ready for a write or read operation. Mount time was calculated for and expressed as follows:

The Cray UNICOS tpmnt command with the -w option was used to control command processing during timed mount testing.

1. Mount time in seconds.

2. Mount rate: total number of mounts per hour.

A 51 MB data file was used for movement of data during the timed write and read operations. The Cray UNICOS tpmnt command with the -w option was needed to define the beginning of a

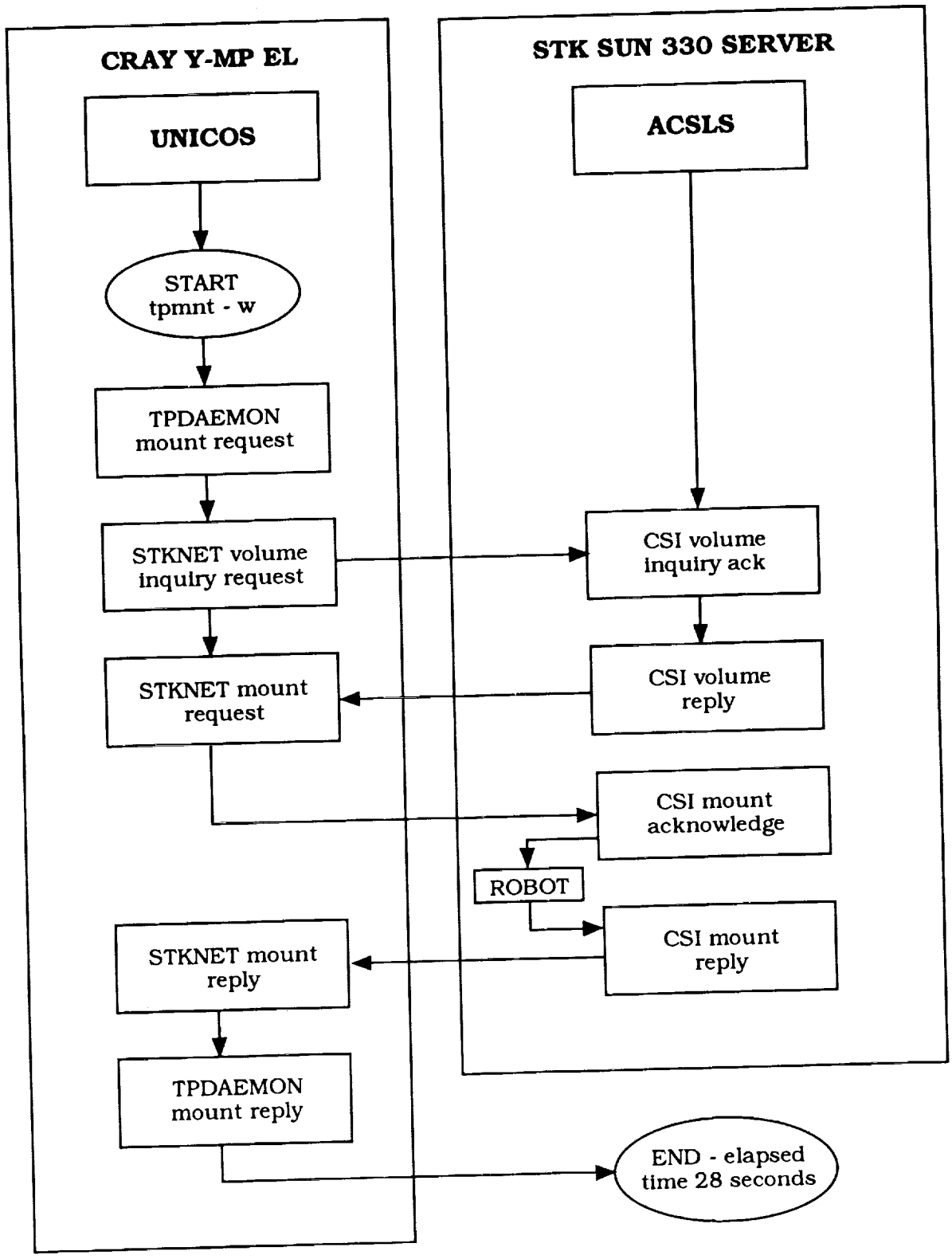

Figure 2. CRAY-STK Sun Server Communications Path (Example - Mount Request)

write/read operation. It was used in conjunction with the Cray UNICOS gettime system command issued before and after the transfer operation to capture actual data transfer time. Transfer rate was calculated and expressed in terms of MB/s.

The transfer rate to tape for a large data file using DMF was also timed. The DMF command, dmput was used with the Cray UNICOS timex command to measure transfer rate for DMF of a 207 MB data file. This measure was expressed in MB/s and included both tape rewind and unload time.

Potential maximum SCSI 1 channel speed was also measured for our configuration. A 207 MB data file was written to tape using the Cray UNICOS tar command with the -b option. The block size specified with the -b option was 128 and corresponds to a 64K byte block size. The tape was read using tar with -b equal to 128 and with the -t option which instructs tar to read only the tape label, but also forces the read process to go to the EOT marker. The time in which tar accomplished this was defined as the potential maximum channel speed for our configuration and was expressed in MB/s.

Test Results

This section summarizes the various test results recorded in this study. Mount times, disk -to -tape read/write transfer rates, tape -to- disk read/write transfer rates, DMF transfer rates, channel speed transfer rates and robotics observations are presented. Average transfer rates are also computed for these functions.

a. Mount Times

It took the Cray Y-MP EL/STK 4400 tape subsystem an average of 52.15 seconds to mount one tape. The Cray Y-MP EL averaged approximately 40-50% user activity during the periods in which the timed mount tests were run. Figures 3, 4 and 5 further break down the results of the timed mount tests by tape device and the controller path to which each was configured. These figures compare mount times for the first and last tape device on each controller, as well as the last tape device on controller 0 to the first tape device on controller 1. Although not significant, the first device configured to each controller averaged slightly lower mount times than the last device configured to the same controller. The average mount times appear to be more a function of position within the controller rather than to the port the controller is configured. The average mount times by controller and tape device are as follows:

	tape device	mount time
controller 0 -	silo00	51.25 seconds
	silo03	52.47 seconds
controller 1	silo04	51.96 seconds
	silo07	52.93 seconds

b. Disk -to -Tape (Tape write operation)

The data transfer rates for disk reads/tape writes are shown in Figures 6 and 7. The block sizes tested are 4KB, 8KB, 16KB, 32KB, 64KB and 128KB. Transparent buffered I/O, a technique which produces tapes with a block size equal to the maximum block size specified by the -b option of the UNICOS tpmnt command was used to move 51 MB's of data from disk to tape in all tests. The last block may vary in size from 1B to the maximum block size. Note that the three smaller block sizes (4KB, 8KB and 16KB) exhibited average transfer rates of less than 1 MB/s while the three larger block sizes (32KB, 64KB and 128KB) exceeded average transfer rates of 1 MB/s. Movement of data to tape in our environment appears to be fastest when using a block size of 64KB. Running these tests consistently caused the test job to use 12% additional system resources regardless of block size. Average tape write operation transfer rates by block size are as follows:

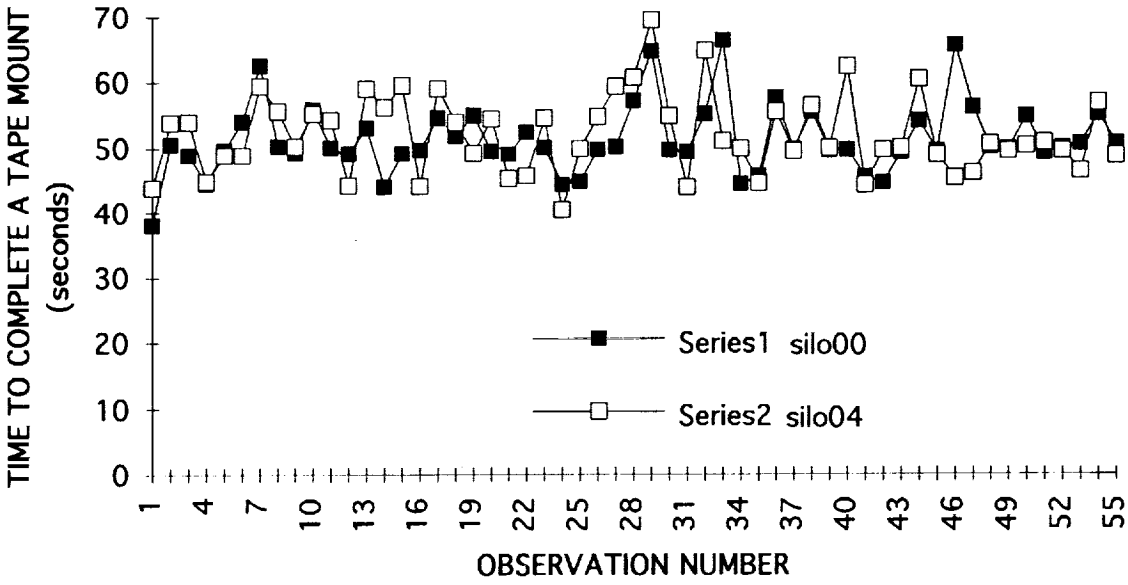

Figure 3. Tape Mounts Silo00 and Silo03 (Timed Mounts for 1 Hour)

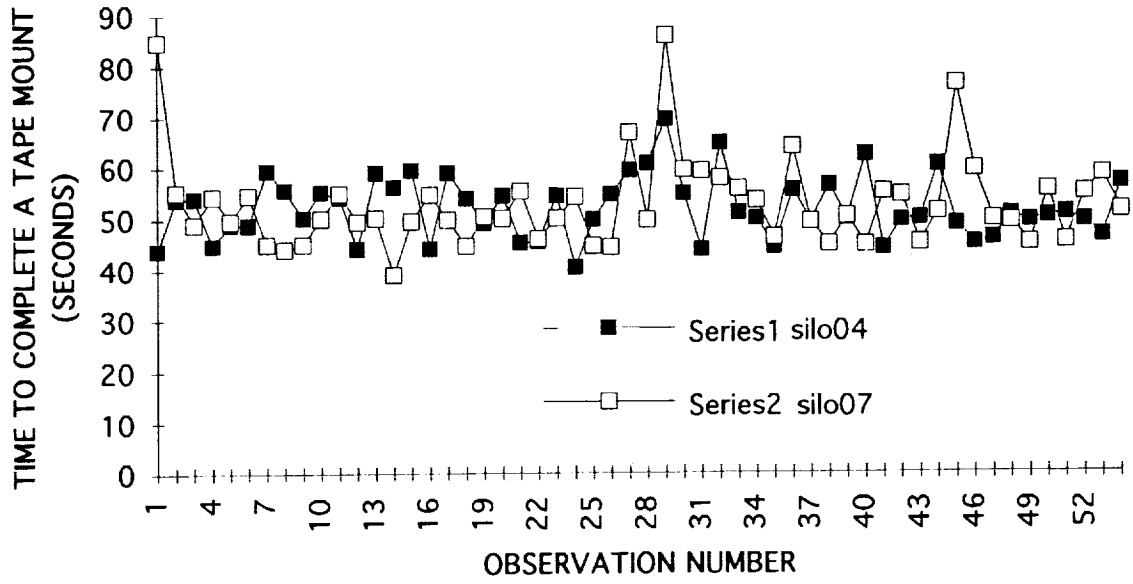

Figure 4. Tape Mount Timings for Silo04 and Silo07 (Average 1 Hour Timed Mounts)

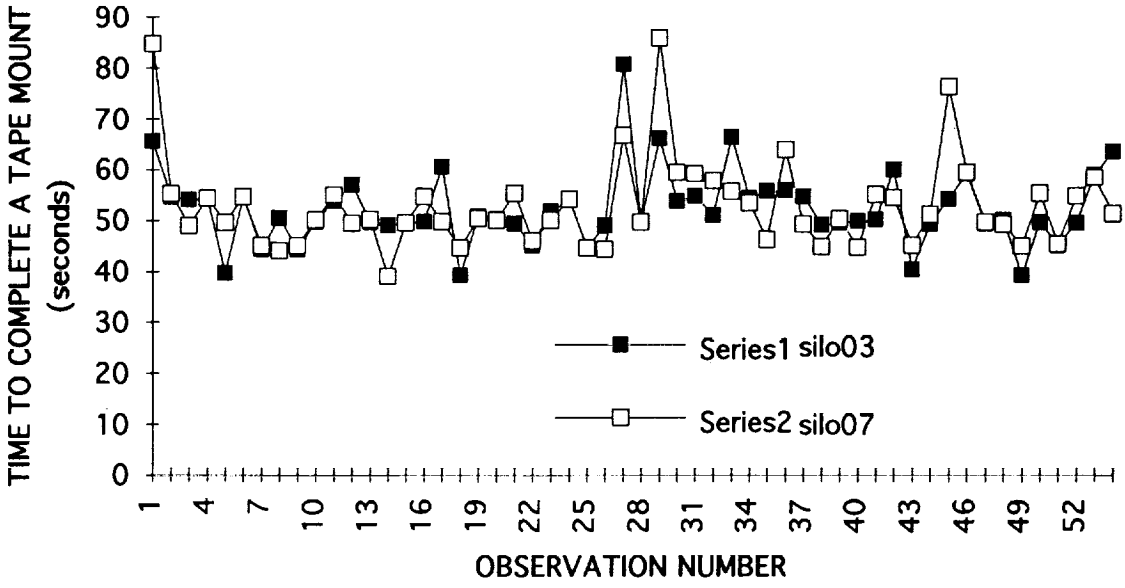

Figure 5. Tape Mount Timings for Silo03 and Silo07 (Averages for 1 Hour Timed Mounts)

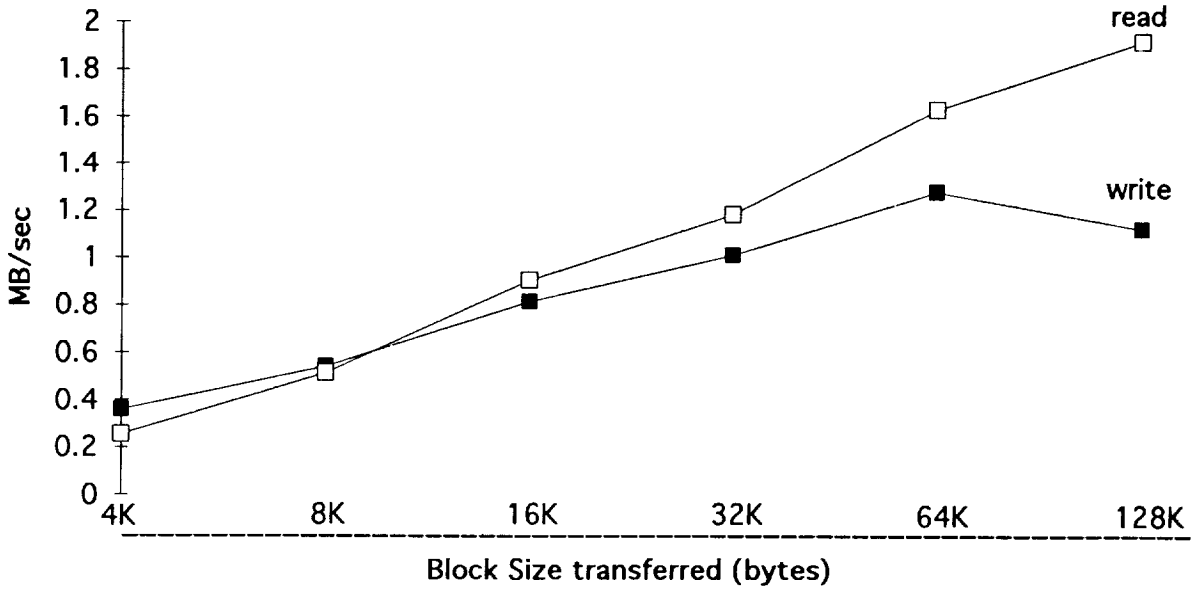

Figure 6. Tape I/O Write/Read Transfer Rates

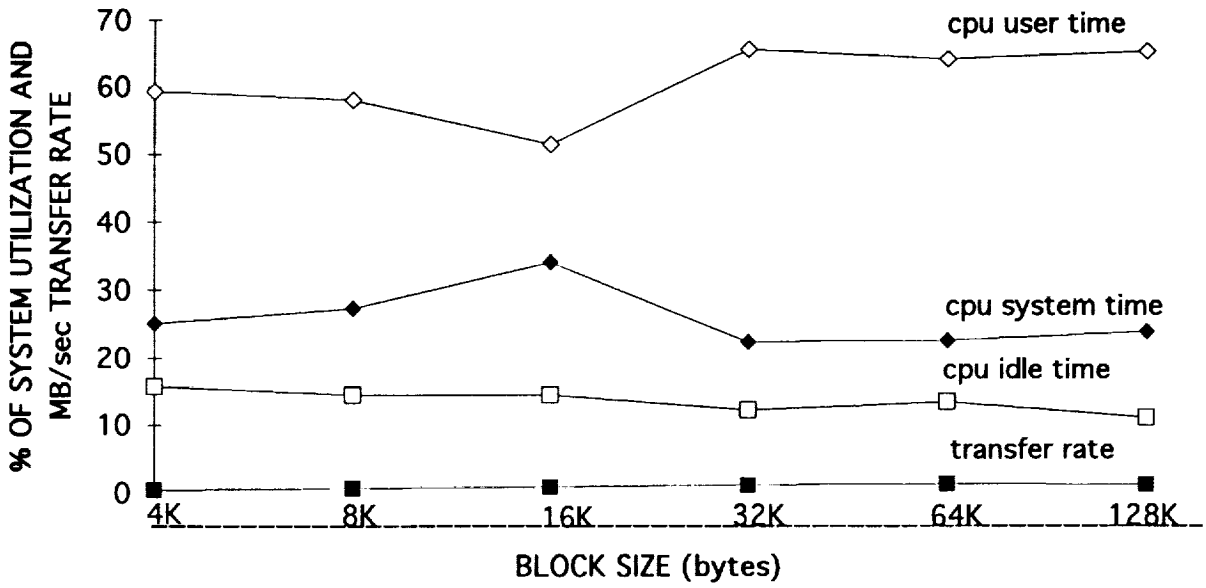

Figure 7. Tape I/O Write and Systems Utilization Time

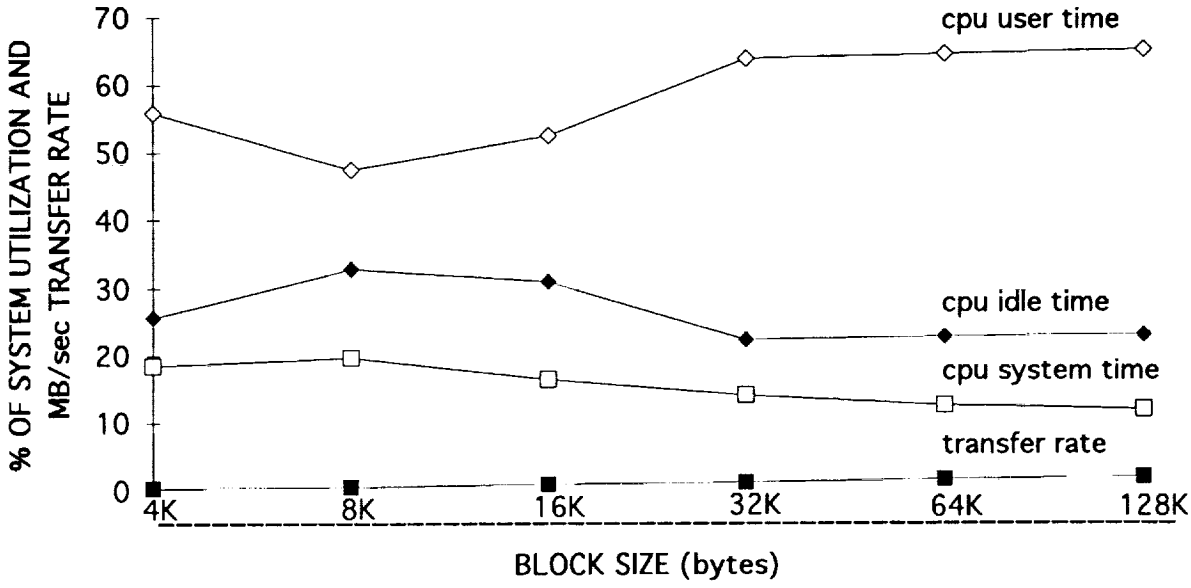

Figure 8. Tape I/O Read and System Utilization Time

Block Size	Transfer Rate
4KB	0.36 MB/s
8KB	0.54 MB/s
16KB	0.91 MB/s
32KB	1.01 MB/s
64KB	1.28 MB/s
128KB	1.13 MB/s

c. Tape -to -Disk (tape read operations)

The data transfer rates for tape reads/disk writes are shown in Figures 6 and 8. The same block sizes used to write a tape were used to read the tape. The technique of transparent buffered I/O was also applied. The three smaller block sizes consistently averaged less than 1 MB/s transfer rate and beginning with 32KB block sizes, the average transfer rate exceeded 1MB/s. For the block sizes tested in our environment, a 128KB block size achieved the fastest transfer rate of 1.92 MB/s. In our configuration, tape read s consistently required 10% more CPU time than would otherwise be required regardless of the blocking factor used. Average tape read operation transfer rates by block size are indicated below:

Block Size	Transfer Rate
4KB	0.26 MB/s
8KB	0.52 MB/s
16KB	0.91 MB/s
32KB	1.19 MB/s
64KB	1.63 MB/s
128KB	1.92 MB/s

d. DMF

The achieved data transfer rate for DMF was 1.39 MB/s. This transfer rate remained consistent regardless of controller and tape device. The time measured included tape volume rewind and unload time.

e. Channel Speed

The measured channel speed for the SCSI 1 ports remained consistent at 2.59 MB/s regardless of controller and device selected. Because this figure included tape volume rewind and unload time we suspect that the SCSI 1 port channel bandwidth of 4.5 MB/s was approached if not realized. However, a measure of tape rewind and unload time was not taken to confirm this.

f. *Robotic Observations*

Considerable effort was taken to observe the motion characteristics of the STK 4400 robotics during the timed tape mount testing. These observations proved invaluable during the design phase of the C programs used to time the tape mounts and contributed to the following design changes:

1. Use the -w option on the UNICOS tpmnt command.
2. Use at least two tape volumes in each timed mount test job.
3. Use tape volumes not housed in the same general area of the STK 4400 ACS library.
4. Write a zero byte tar file to the tapes used for timing mounts and read them during the test phase by issuing the UNICOS tar command with the -t option.

These decisions eliminated the inherent bias imposed by the design of the STK 4400 hardware and firmware.

The robot would remain "at home" in the position it was last instructed to go. Using only one tape volume would always put the robot in front of that tape regardless of location within the library.

When the Cray Y-MP EL executed a tpmnt request without the -w option it would immediately begin processing the next command without waiting for acknowledgment from the STK Sun 330 server that the tape was in fact ready for a read or write operation. If the next command was a UNICOS rls -a (to release all resources) the robot would never physically mount the tape. It would sit in front of the tape volume still in its cell location within the library and never complete the fetch operation.

Once the design changes were implemented,, as many as four timed mount sessions were able to be run simultaneously, using eight different tape volumes. The intelligence built into the STK 4400 robotics system did not then appear to bias or impact our results.

Discussion

a. Hardware Considerations

The transfer rate for tape I/O depends on several factors, the most important of which is the block size used for each transfer. However, to a much larger extent than would be evident at first sight, this transfer rate also depends on the characteristics of the hardware being used.

The I/O capabilities of the Cray Y-MP EL is impacted by its hardware design. The IOBB (see Figure 1) is designed to support a maximum block size of 128KB. In addition, the IOBB must be used to support all I/O transfers for all peripheral devices configured on the Cray Y-MP EL system. This is a limitation which cannot be modified by a user. All simultaneous I/O operations, tape and disk alike, compete for IOBB resources. In our configuration, this built in contention was exacerbated by the lack of available disk space (4 disk drives configured, see Figure 1). One job writing to tape while at the same time another job is trying to read from tape, will always force the second job to be put in a wait for I/O state until the first job completes. This wait time accumulates as part of the transfer time and results in a slower overall transfer rate for the second job.

Some of this contention can be "programmed out" by insuring that the file (i.e. being written to tape) resides on a file system that is not heavily used. We did not use the /tmp file system for this reason.

In our configuration, both tape controllers were attached to the same Cray Y-MP EL IOS and used the same IOBB. Contention for tape resources were evident in multiple session tape jobs and exhibited by both higher mount times and lower transfer rates. In a single session mount job we were able to achieve an average mount time of 28 seconds, but while running multiple mount jobs this time approached 52 seconds.

The Ciprico SCSI controller, with the IOBB does not permit block sizes larger than 128 KB. While the read transfer rate increased progressively for block sizes up to 128K, the write rate actually *decreased* for the same block size. We did not expect the disk -to -tape transfer rate to peak out at the 64K block size, since Cray supports block sizes of up to 4 MB in Unicos 6.1.6. The buffering in the 4781 caused the write rate to drop because the buffer set-up time increased the amount of time required to transfer data. The 4781 has a total of 1 MB of buffer space which is shared equally among all drives configured. Since we have eight tape drives we had only 125KB of buffer space allocated for each tape device. A 128K block size caused a delay in write operations, due to preparing the 125KB buffer space. In read operations, the transparent buffered I/O transfer technique worked well and allowed us to successfully stream our data.

b. Robotic Considerations

During this project we uncovered several important factors surrounding the motion and responsiveness of the robot. Our goal was not only to measure the time for various operations, but also to characterize the performance of the robot under different operational conditions.

Immediately noted was the fact that the robot would always find as its home position, the location in front of the last cartridge serviced (i.e., on a tape mount, it would remain in front of the tape drive; on a release, it would remain in front of the tape's cell location). Such a home location would only change once a new command to mount or dismount a different tape was received.

c. UNICOS Issues

During the testing, two significant problems related to UNICOS on the Cray Y-MP EL were noted with consistency. The first involved crashing the system when trying to read a previously written 1K block. This problem was attributed to Unicos internals and promptly resolved by Cray.

The second problem involved the tape daemon. The tape daemon hung consistently while trying concurrent tape I/O with four sessions by the same user. This problem was not resolved and is currently under investigation.

Another aspect of library usage also involved the tape daemon and its technique of allocating drives to user jobs. The Unicos tape daemon assigns tape drives in a round-robin fashion and does not take into consideration the location of a requested volume when assigning a drive to this volume. In library configurations of multiple silos this can have a serious impact on efficiency. Given the tape daemon's round-robin policy of assigning drives, a volume could be assigned a drive not attached to the silo in which that volume resides. The net effect would be to at least double the effective mount and dismount times because of "pass-throughs" that would be required to complete the tape request.

f. *Software Considerations*

Overall performance as perceived by a user is dependent upon the storage and file management system used to store data in the library. Cray's Data Migration Facility was used to manage data stored in the STK ACS 4400 library. An additional goal of this project was to measure and characterize DMF. We were able to achieve a very respectable transfer rate of 1.39 MB/s. Unfortunately, the ACS 4400 library became unavailable to us in the test environment shortly after completing this phase of our testing and we were unable to further examine DMF as a file management system.

With DMF we anticipated faster transfer rates due to its larger internal block size (49 KB) as compared to other popular file management systems, such as UniTree. Some versions of UniTree use a 15.5KB blocking factor, which may result in slower transfer times.

Summary

This study of the STK ACS 4400 Tape library revealed important information concerning on-line mass storage space. While the STK Tape Library performed well, it did not achieve the manufacturer's advertised specifications in our test environment. We see channel port type as the primary limiting factor. Maximum throughput could be enhanced by upgrading to a SCSI 2 port, if available, or to a Block Mux port, which was used by the manufacturer in a highly controlled IBM test environment to achieve the figures they report.

Transfer rates from our study, which more closely emulate a real user environment, showed a direct correlation with block size. Other constraints, primarily due to inherent hardware

limitations, were circumvented by code modifications that optimized system commands and by using alternate file systems to reduce I/O contention.

The speed at which data can be transferred is affected by both hardware and software considerations as shown in our study. However, by applying the techniques reported in this paper, the requirement for reliable movement of data from disk to tape and from tape to disk would certainly be achieved.

Acknowledgements

The authors appreciate the numerous helpful questions, comments and suggestions from Ben Kobler and Dr. P.C. Hariharan. Ray Yee and Frithjov Iverson of Cray Research gave valuable assistance in identifying and understanding specific UNICOS I/O features. Steve Cranage of STK offered useful insights into the ACS 4400 Library operation.

References

1. S. Ranade, Mass Storage Technologies (Meckler Corp. Westport, CT 1990)

2. UNICOS File Formats and Special Files Reference Manual, Publication SR2014, Cray Research Inc.

3. UNICOS User Commands Reference Manual, Publication SR2011, Cray Research Inc.